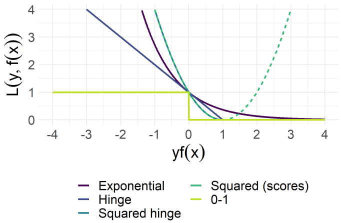
Introduction to Machine Learning

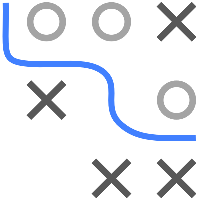
Advanced Classification Losses



Learning goals

- Know the (squared) hinge loss
- Know the *L*2 loss defined on scores
- Know the exponential loss
- Know the AUC loss

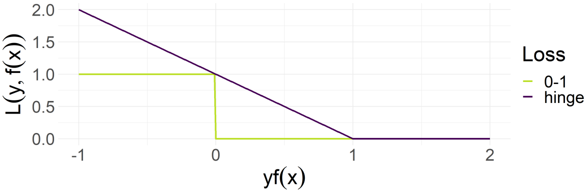
HINGE LOSS



- The intuitive appeal of the 0-1-loss is set off by its analytical properties ill-suited to direct optimization.
- The **hinge loss** is a continuous relaxation that acts as a convex upper bound on the 0-1-loss (for $y \in \{-1, +1\}$):

$$L(y, f(\mathbf{x})) = \max\{0, 1 - yf(\mathbf{x})\}.$$

- Note that the hinge loss only equals zero for a margin ≥ 1, encouraging confident (correct) predictions.
- It resembles a door hinge, hence the name:

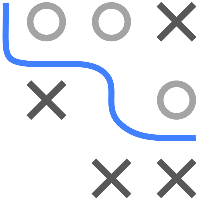


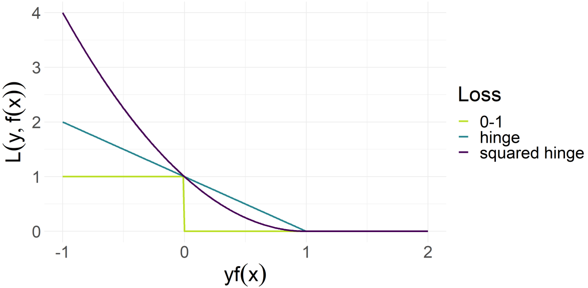
SQUARED HINGE LOSS

- We can also specify a **squared** version for the hinge loss:

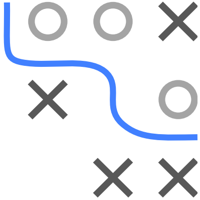
$$L(y, f(\mathbf{x})) = \max\{0, (1 - yf(\mathbf{x}))\}^2.$$

- The $L2$ form punishes margins $yf(\mathbf{x}) \in (0, 1)$ less severely but puts a high penalty on more confidently wrong predictions.
- Therefore, it is smoother yet more outlier-sensitive than the non-squared hinge loss.





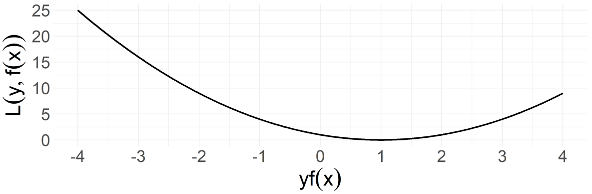
SQUARED LOSS ON SCORES



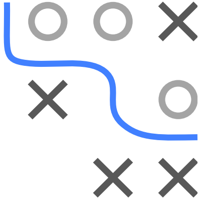
- Analogous to the Brier score defined on probabilities we can specify a **squared loss on classification scores** (again, $y \in \{-1, +1\}$, using that $y^2 \equiv 1$):

$$\begin{aligned} L(y, f(\mathbf{x})) = & \quad (y - f(\mathbf{x}))^2 = y^2 - 2yf(\mathbf{x}) + (f(\mathbf{x}))^2 = \\ = & \quad 1 - 2yf(\mathbf{x}) + (yf(\mathbf{x}))^2 = (1 - yf(\mathbf{x}))^2 \end{aligned}$$

- This loss behaves just like the squared hinge loss for $yf(\mathbf{x}) < 1$, but is zero only for $yf(\mathbf{x}) = 1$ and actually increases again for larger margins (which is in general not desirable!)

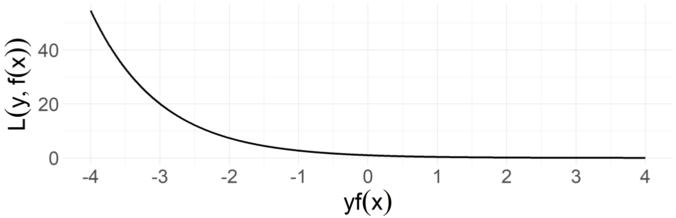


CLASSIFICATION LOSSES: EXPONENTIAL LOSS

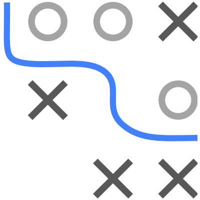


Another possible choice for a (binary) loss function that is a smooth approximation to the 0-1-loss is the **exponential loss**:

- $L(y, f(\mathbf{x})) = \exp(-yf(\mathbf{x}))$, used in AdaBoost.
- Convex, differentiable (thus easier to optimize than 0-1-loss).
- The loss increases exponentially for wrong predictions with high confidence; if the prediction is right with a small confidence only, there, loss is still positive.
- No closed-form analytic solution to (empirical) risk minimization.



CLASSIFICATION LOSSES: AUC-LOSS



- Often AUC is used as an evaluation criterion for binary classifiers.
- Let $y \in \{-1, +1\}$ with n_- negative and n_+ positive samples.
- The AUC can then be defined as

$$AUC = \frac{1}{n_+} \frac{1}{n_-} \sum_{i: y^{(i)} = 1} \sum_{j: y^{(j)} = -1} [f^{(i)} > f^{(j)}]$$

- This is not differentiable w.r.t f due to $[f^{(i)} > f^{(j)}]$.
- But the indicator function can be approximated by the distribution function of the triangular distribution on $[-1, 1]$ with mean 0.
- However, direct optimization of the AUC is numerically more difficult, and might not work as well as using a common loss and tuning for AUC in practice.